We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

1

Bio-Inspired Active Vision Paradigms in Surveillance Applications

Mauricio Vanegas, Manuela Chessa, Fabio Solari and Silvio Sabatini
The Physical Structure of Perception and Computation - Group, University of Genoa
Italy

1. Introduction

Visual perception was described by Marr (1982) as the processing of visual stimuli through three hierarchical levels of computation. In the first level or *low-level* vision it is performed the extraction of fundamental components of the observed scene such as edges, corners, flow vectors and binocular disparity. In the second level or *medium-level* vision it is performed the recognition of objects (e.g. model matching and tracking). Finally, in the third level or *high-level* vision it is performed the interpretation of the scene. A complementary view is presented in (Ratha & Jain, 1999; Weems, 1991); by contrast, the processing of visual stimuli is analysed under the perspective developed by Marr (1982) but emphasising how much data is being processed and what is the complexity of the operators used at each level. Hence, the low-level vision is characterised by large amount of data, small neighbourhood data access, and simple operators; the medium-level vision is characterised by small neighbourhood data access, reduced amount of data, and complex operators; and the high-level vision is defined by non-local data access, small amount of data, and complex relational algorithms. Bearing in mind the different processing levels and their specific characteristics, it is plausible to describe a computer vision system as a modular framework in which the low-level vision processes can be implemented by using parallel processing engines like GPUs and FPGAs to exploit the data locality and the simple algorithmic operations of the models; and the medium and high-level vision processes can be implemented by using CPUs in order to take full advantage of the straightforward fashion of programming these kind of devices.

The low-level vision tasks are probably the most studied in computer vision and they are still an open research area for a great variety of well defined problems. In particular, the estimation of optic flow and of binocular disparity have earned special attention because of their applicability in segmentation and tracking. On the one hand, the stereo information has been proposed as a useful cue to overcome some of the issues inherent to robust pedestrian detection (Zhao & Thorpe, 2000), to segment the foreground from background layers (Kolmogorov et al., 2005), and to perform tracking (Harville, 2004). On the other hand, the optic flow is commonly used as a robust feature in motion-based segmentation and tracking (Andrade et al., 2006; Yilmaz et al., 2006).

This chapter aims to describe a biological inspired video processing system for being used in video surveillance applications; the degree of similarity between the proposed framework

and the human visual system allows us to take full advantage of both optic flow and disparity estimations not only for tracking and fixation in depth but also for scene segmentation. The most relevant aspect in the proposed framework is its hardware and software modularity. The proposed system integrates three cameras (see Fig. 1); two active cameras with variable-focal-length lenses (binocular system) and a third fixed camera with a wide-angle lens. This system has been designed to be compatible with the well-known iCub robot interface[1]. The cameras movement control, as well as the zoom and iris control run on an embedded computer PC/104. The optic flow and the disparity algorithms run on a desktop computer equipped with a processor *Intel Core 2 Quad @ 2.40GHz* and a memory RAM of about 8 GB. All system components, namely the desktop computer, the embedded computer PC/104, and the cameras, are connected in a gigabit Ethernet network through which they can interact as a distributed system.

Fig. 1. Trinocular robotic head with 5 degrees of freedom, namely a common tilt movement, and independent zoom-pan movements for left and right cameras, respectively.

The general features of the moving platform are compiled in Table 1. Likewise, the optic features of the cameras are collected in Table 2. Lastly, it is important to mention that the binocular system has a baseline of 30 cm.

Features	**Pan Movement**	**Tilt Movement**
Limits:	$\pm 30^\circ$ (Software limit)	$\pm 60^\circ$ (Software limit)
Acceleration:	$5100^\circ/sec^2$	$2100^\circ/sec^2$
Max. Speed:	$330^\circ/sec$	$73^\circ/sec$
Resolution:	0.03°	0.007°
Optical Encoder:	512 pulses/revolution	512 pulses/revolution
Motor Voltage:	12 V	12 V
Gear Ratio:	1:80	1:80
Motor Torque:	0.59 Nm	0.59 Nm

Table 1. General features of the moving platform.

Most of the video surveillance systems are networks of cameras for a proper coverage of wide areas. These networks use both fixed or active cameras, or even a combination of both, placed

[1] The iCub is the humanoid robot developed as part of the EU project RobotCub and subsequently adopted by more than 20 laboratories worldwide (see `http://www.icub.org/`).

Features	Active Cameras	Fixed Camera
Resolution:	11392 x 1040 pixels	1624 x 1236 pixels
Sensor Area:	6.4 x 4.8 mm	7.1 x 5.4 mm
Pixel Size:	4.65 x 4.65 μm	4.4 x 4.4 μm
Focal Length:	7.3 $\sim$ 117 mm, FOV 47° $\sim$ 3°	4.8 mm, FOV 73°

Table 2. Optic features of the cameras.

at not predetermined positions to strategically cover a wide area; the term *active* specifies the camera's ability of changing both the angular position and the field of view. The type of cameras used in the network has inspired different calibration processes to find automatically both the intrinsic and extrinsic camera parameters. In this regard, Lee et al. (2000) proposed a method to estimate the 3D positions and orientations of fixed cameras, and the ground plane in a global reference frame which lets the multiple cameras views to be aligned into a single planar coordinate frame; this method assume approximate values for intrinsic cameras parameters and it is based on overlapped cameras views; however, others calibration methods have been proposed for non-overlapped cameras views (i.e. Kumar et al., 2008). In the case of active cameras, Tsai (1987) has developed a method for estimating both the matrices of rotation and translation in the Cartesian reference frame, and the intrinsic parameters of the cameras. In addition to the calibration methods, the current surveillance systems must deal with the segmentation and identification of complex scenes in order to characterise them and thus to obtain a classification which let the system to recognise unusual behaviours into the scene. In this regard, a large variety of algorithms have been developed to detect changes in scene; for example the application of a threshold to the absolute difference between pixel intensities of two consecutive frames can lead to the identification of moving objects, some methods for the threshold selection are described in (Kapur et al., 1985; Otsu, 1979; Ridler & Calvar, 1978). Other examples are the adaptive background subtraction to detect moving foreground objects (Stauffer & Grimson, 1999; 2000) and the estimation of optic flow (Barron et al., 1994). Our proposal differs the most of the current surveillance systems in at least three aspects: (1) the use of a single camera with a wide-angle lens to cover vast areas and a binocular system for tracking areas of interest at different fields of view (the wide-angle camera is used as the reference frame), (2) the estimation of both optic flow and binocular disparity for segmenting the images; this system feature can provide useful information for disambiguating occlusions in dynamic scenarios, and (3) the use of a bio-inspired fixation strategy which lets the system to fixate areas of interest, accurately.

In order to explain the system behaviour, two different perspectives were described. On the one hand, we present the system as a bio-inspired mathematical model of the primary visual cortex (see section 2); from this viewpoint, we developed a low-level vision architecture for estimating optic flow and binocular disparity. On the other hand, we describe the geometry of the cameras position in order to derive the equations that govern the movement of the cameras (see section 3). Once the system is completely described, we define an angular-position control capable of changing the viewpoint of the binocular system by using disparity measures in section 4. An interesting case study is described in section 5 where both disparity and optic flow are used to segment images. Finally, in section 6, we present and discuss the system's performance results.

2. The system: a low-level vision approach

The visual cortex is the largest, and probably the most studied part of the human brain. The visual cortex is responsible for the processing of visual stimuli impinging on the retinas. As a matter of fact, the first stage of processing takes place in the lateral geniculate nucleus (LGN) and then the neurons of the LGN relay the visual information to the primary visual cortex (V1). Then, the visual information flow hierarchically to areas V2, V3, V4 and V5/MT where visual perception gradually takes place.

The experiments carried out by Hubel & Wiesel (1968) proved that the primary visual cortex (V1) consists of cells responsive to different kinds of spatiotemporal features of the visual information. The apparent complexity with which the brain extracts the spatiotemporal features has been clearly explained by Adelson & Bergen (1991). The light filling a region of space contains information about the objects in that space; in this regard, they proposed the *plenoptic function* to describe mathematically the pattern of light rays collected by a vision system. By definition, the plenoptic function describes the state of luminous environment, thus the task of the visual system is to extract structural elements from it.

Structural elements of the plenoptic function can be described as oriented patterns in the plenoptic space, and the primary cortex can be interpreted as a set of local, Fourier or Gabor operators used to characterise the plenoptic function in the spatiotemporal and frequency domains.

2.1 Neuromorphic paradigms for visual processing

Mathematically speaking, the extraction of the most important aspects of the plenoptic function can emulate perfectly the neuronal processing of the primary visual cortex (V1). More precisely, qualities or elements of the visual input can be estimated by applying a set of low order directional derivatives at the sample points; the so obtained measures represent the amount of a particular type of local structure. To effectively characterise a function within a neighbourhood, it is necessary to work with the local average derivative or, in an equivalent form, with the oriented linear filters in the function hyperplanes. Consequently, the neurons in V1 can be interpreted as a set of oriented linear filters whose outputs can be combined to obtain more complex feature detectors or, what is the same, more complex receptive fields. The combination of linear filters allow us to measure the magnitude of local changes within a specific region, without specifying the exact location or spatial structure. The receptive fields of complex neurons have been modelled as the sum of the squared responses of two linear receptive fields that differ just in phase for 90° (Adelson & Bergen, 1985); as a result, the receptive fields of complex cells provide *local energy measures*.

2.2 Neural Architecture to estimate optic flow and binocular disparity

The combination of receptive fields oriented in space-time can be used to compute local energy measures for optic flow (Adelson & Bergen, 1985). Analogously, by combining the outputs of spatial receptive fields it is possible to compute local energy measures for binocular disparity (Fleet et al., 1996; Ohzawa et al., 1990). On this ground, it has been recently proposed a neural architecture for the computation of horizontal and vertical disparities and optic flow (Chessa, Sabatini & Solari, 2009). Structurally, the architecture comprises four processing stages (see

Fig. 2): the distributed coding of the features by means of oriented filters that resemble the filtering process in area V1; the decoding process of the filter responses; the estimation of the local energy for both optic flow and binocular disparity; and the coarse-to-fine refinement.

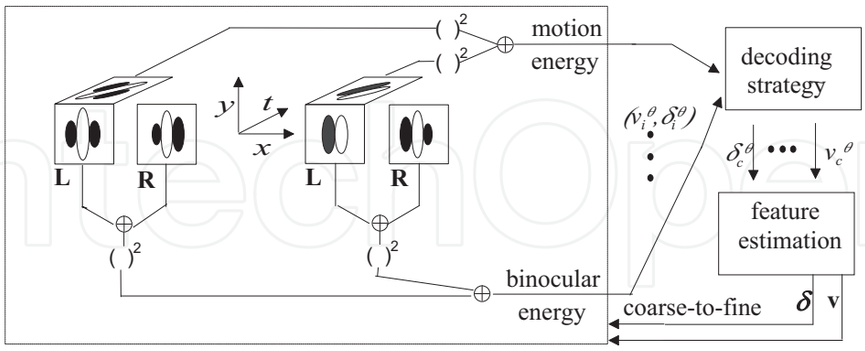


Fig. 2. The neural architecture for the computation of disparity and optic flow.

The neuronal population is composed of a set of 3D Gabor filters which are capable of uniformly covering the different spatial orientations, and of optimally sampling the spatiotemporal domain (Daugman, 1985). The linear derivative-like computation concept of the Gabor filters let the filters to have the form $h(\mathbf{x},t) = g(\mathbf{x})f(t)$. Both spatial and temporal terms in the right term are comprised of one harmonic function and one Gaussian function. This can be easily deduced from the impulse response of the Gabor filter.

The mathematical expression of the spatial term of a 3D Gabor filter rotated by an angle θ with respect to the horizontal axis is:

$$g(x,y;\psi,\theta) = e^{\left(-\frac{x_\theta^2}{2\sigma_x^2}-\frac{y_\theta^2}{2\sigma_y^2}\right)} e^{j(\omega_0 x_\theta+\psi)}, \tag{1}$$

where $\theta \in [0, 2\pi)$ represents the spatial orientation; ω_0 and ψ are the frequency and phase of the sinusoidal modulation, respectively; the values σ_x and σ_y determine the spatial area of the filter; and (x_θ, y_θ) are the rotated spatial coordinates.

The algorithm to estimate the binocular disparity is based on a *phase-shift* model; one of the variations of this model suggests that disparity is coded by phase shifts between receptive fields of the left and right eyes whose centres are in the same retinal position (Ohzawa et al., 1990). Let the left and right receptive fields be $g^L(\mathbf{x})$ and $g^R(\mathbf{x})$, respectively; the binocular phase shift is defined by $\Delta\psi = \psi^L - \psi^R$. Each spatial orientation has a set of k receptive fields with different binocular phase shifts in order to be sensitive to different disparities ($\delta^\theta = \Delta\psi/\omega_o$); the phase shifts are uniformly distributed between $-\pi$ and π. Therefore, the left and right receptive fields are applied to a binocular image pair $I^L(\mathbf{x})$ and $I^R(\mathbf{x})$ according to the following equation:

$$Q(\mathbf{x_0};\delta^\theta) = \int_{-\infty}^{\infty} g^L(\mathbf{x_0}-\mathbf{x}) I^L(\mathbf{x}) d\mathbf{x} + \int_{-\infty}^{\infty} g^R(\mathbf{x_0}-\mathbf{x}) I^R(\mathbf{x}) d\mathbf{x}, \tag{2}$$

so, the spatial array of binocular energy measures can be expressed as:

$$E(\mathbf{x};\delta^\theta) = |Q(\mathbf{x};\delta^\theta)|^2 = |Q^L(\mathbf{x};\delta^\theta) + e^{-j\Delta\psi} Q^R(\mathbf{x};\delta^\theta)|^2. \tag{3}$$

Likewise, the temporal term of a 3D Gabor filter is defined by:

$$f(t;\omega_t) = e^{\left(-\frac{t^2}{2\sigma_t^2}\right)} e^{j\omega_t t} 1(t), \tag{4}$$

where σ_t determines the integration window of the filter in time domain; ω_t is the frequency of the sinusoidal modulation; and $1(t)$ denotes the unit step function. Each receptive field is tuned to a specific velocity v^θ along the direction orthogonal to the spatial orientation θ. The temporal frequency is varied according to $\omega_t = v^\theta \omega_0$. Each spatial orientation has a set of receptive fields sensitive to M tuning velocities; M depends on the size of the area covered by each filter according to the Nyquist criterion.

The set of spatiotemporal receptive fields $h(\mathbf{x}, t)$ is applied to an images sequence $I(\mathbf{x}, t)$ according to the following equation:

$$Q(\mathbf{x_0}, t; v^\theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\mathbf{x_0} - \mathbf{x}, t - \tau) I(\mathbf{x}, \tau) d\mathbf{x} d\tau, \tag{5}$$

so, the motion energy $E(\mathbf{x_0}, t; v^\theta)$ equals:

$$E(\mathbf{x_0}, t; v^\theta) = |Q(\mathbf{x_0}, t; v^\theta)|^2 = \left| e^{j\psi(t)} \int_0^t Q(\mathbf{x_0}, \tau; v^\theta) e^{-j\omega_t \tau} d\tau \right|^2. \tag{6}$$

where $\psi(t) = \psi + \omega_t t = \psi + \omega_0 v^\theta t$.

So far, we have described the process of encoding both binocular disparity and optic flow by means of a $N \times M \times K$ array of filters uniformly distributed in space domain. Now, it is necessary to extract the component velocity (v_c^θ) and the component disparity (δ_c^θ) from the local energy measures at each spatial orientation. The accuracy in the extraction of these components is strictly correlated with the number of filters used per orientation, such that precise estimations require a large number of filters; as a consequence, it is of primary importance to establish a compromise between the desired accuracy and the number of filters used or, what is the same, a compromise between accuracy and computational cost.

An affordable computational cost can be achieved by using weighted sum methods as the *maximum likelihood* proposed by Pouget et al. (2003). However, the proposed architecture uses the centre of gravity of the population activity since it has shown the best compromise between simplicity, computational cost and reliability of the estimates. Therefore, the component velocity v_c^θ is obtained by pooling cell responses over all orientations:

$$v_c^\theta(\mathbf{x_0}, t) = \frac{\sum_{i=1}^{M} v_i^\theta E(\mathbf{x_0}, t; v_i^\theta)}{\sum_{i=1}^{M} E(\mathbf{x_0}, t; v_i^\theta)}, \tag{7}$$

where v_i^θ represent all the M tuning velocities; and $E(\mathbf{x_0}, t; v_i^\theta)$ represent the motion energies at each spatial orientation. The component disparity δ_c^θ can be estimated in a similar way.

Because of the aperture problem a filter can just estimate the features which are orthogonal to the orientation of the filter. So we adopt k different binocular and M different motion receptive fields for each spatial orientation; consequently, a robust estimate for the full velocity $\mathbf{v}$ and for

the full disparity $\boldsymbol{\delta}$ is achieved by combining all the estimates v_c^θ and δ_c^θ, respectively (Pauwels & Van Hulle, 2006; Theimer & Mallot, 1994).

Finally, the neural architecture uses a coarse to fine control strategy in order to increase the range of detection in both motion and disparity. The displacement features obtained at coarser levels are expanded and used to warp the images in finer levels in order to achieve a higher displacement resolution.

3. The system: a geometrical description

In the previous section we presented the system from a biological point of view. We have summarised a mathematical model of the behaviour of the primary visual cortex and we have proposed a computational architecture based on linear filters for estimating optic flow and binocular disparity. Now it is necessary to analyse the system from a geometrical point of view in order to link the visual perception to the camera movements, thus letting the system to interact with the environment.

To facilitate the reference to the cameras within this text, we are going to refer the fixed camera as *wide-angle camera*, and the cameras of the binocular system as *active cameras*. The wide-angle camera is used for a wide view of the scene, and it becomes the reference of the system. In vision research, the cyclopean point is considered the most natural centre of a binocular system (Helmholtz, 1925) and it is used to characterise stereopsis in human vision (Hansard & Horaud, 2008; Koenderink & van Doorn, 1976). By doing a similar approximation, the three-camera model uses the wide-angle-camera image as the cyclopean image of the system. In this regard, the problem statement is not trying to construct the cyclopean image from the binocular system, but using the third camera image as a reference coordinate to properly move the active cameras according to potential targets or regions of interest in the wide range scenario.

Each variable-focal-length camera can be seen as a 3DOFs pan-tilt-zoom (PTZ) camera. However, the three-camera system constraints the active cameras to share the tilt movement due to the mechanical design of the binocular framework. One of the purposes of our work is to describe the geometry of the three-camera system in order to properly move the pan-tilt-zoom cameras to fixate any object in the field of view of the wide-angle camera and thus to get both a magnified view of the target object and the depth of the scene.

We used three coordinates systems to describe the relative motion of the active cameras with respect to the wide-angle camera (see Fig. 3). The origin of each coordinate system is supposed in the focal point of each camera and the Z-axes are aligned with the optical axes of the cameras. The pan angles are measured with respect to the planes $X_L = 0$ and $X_R = 0$ respectively; note that pan angles are positive for points to the left of these planes ($X_L > 0$ or $X_R > 0$). The rotation axes for the pan movement are supposed to be parallel. The common tilt angle is measured with respect to the horizontal plane; note that the tilt angle is positive for points above the horizontal plane ($Y_L = Y_R > 0$).

The point $P(X, Y, Z)$ can be written in terms of the coordinate systems shown in Fig. 3 as follows:

$$(X, Y, Z) = (X_L, Y_L, Z_L) - O_L, \tag{8}$$

$$(X, Y, Z) = (X_R, Y_R, Z_R) - O_R, \tag{9}$$

where $O_L = (dx_L, dy_L, dz_L)$ and $O_R = (-dx_R, dy_R, dz_R)$ are the origin of the coordinate system of the left and right cameras with respect to the wide-angle camera coordinate system.

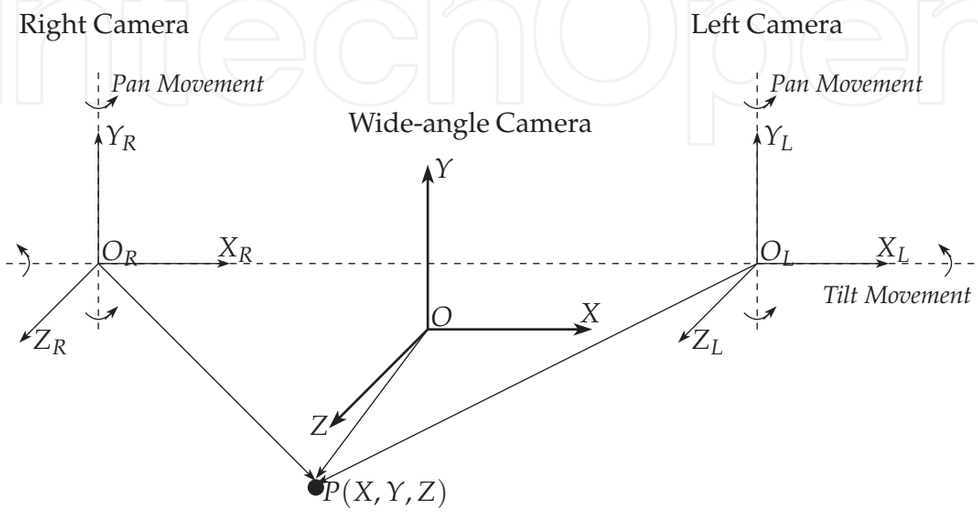


Fig. 3. The coordinate systems of the three cameras in the binocular robotic head.

It is considered f_w as the focal length of the wide-angle camera and f as the focal length of the active cameras. The Equations 8 and 9 can be written in terms of the image coordinate system of the wide-angle camera if these equations are multiplied by factor $\frac{f_w}{Z}$:

$$\frac{f_w}{Z}(X_L, Y_L, Z_L) = (x, y, f_w) + \frac{f_w}{Z}(dx_L, dy_L, dz_L), \tag{10}$$

$$\frac{f_w}{Z}(X_R, Y_R, Z_R) = (x, y, f_w) + \frac{f_w}{Z}(-dx_R, dy_R, dz_R). \tag{11}$$

Now, it is possible to link the image coordinate system of the wide-angle camera to the image coordinate system of the active cameras by multiplying the Equations 10 and 11 by the factors $\frac{f}{Z_L}$ and $\frac{f}{Z_R}$, respectively:

$$\frac{f_w}{Z}(x_L, y_L, f) = \frac{f}{Z_L}(x, y, f_w) + \frac{f_w f}{Z_L Z}(dx_L, dy_L, dz_L), \tag{12}$$

$$\frac{f_w}{Z}(x_R, y_R, f) = \frac{f}{Z_R}(x, y, f_w) + \frac{f_w f}{Z_R Z}(-dx_R, dy_R, dz_R). \tag{13}$$

Assuming that the position of the origin with respect to the Z-axis is small enough compared to the distance of the real object in the scene, it can be done the next approximation $Z \approx Z_L$ and $Z \approx Z_R$. Accordingly, the Equations 12 and 13 can be rewritten to obtain the *wide-to-active*

camera mapping equations as follows:

$$(x_L, y_L, f) = \frac{f}{f_w}(x, y, f_w) + \frac{f}{Z}(dx_L, dy_L, 0), \tag{14}$$

$$(x_R, y_R, f) = \frac{f}{f_w}(x, y, f_w) + \frac{f}{Z}(-dx_R, dy_R, 0). \tag{15}$$

These equations describe the position of any point in the field of view of the wide-angle camera into the image coordinate of the active cameras.

So far, we have described the geometry of the cameras system, now the problem is to transform the wide-to-active camera mapping equations to motor stimuli in order to fixate any point in the wide-angle image. The fixation problem can be defined as the computation of the correct angular position of the motors in charge of the pan and tilt movements of the active cameras, to direct the gaze to any point in the wide-angle image. In this sense, the fixation problem is solved when the point $p(x, y)$ in the wide-angle image can be seen in the centres of the left and right camera images.

From the geometry of the trinocular head we can consider $dx_L = dx_R$, and $dy_L = dy_R$. In this way, both pan (θ_L, θ_R) and tilt (θ_y) angles of the active cameras, according to the wide-to-active camera mapping equations, can be written as:

$$\theta_L = \arctan\left(\frac{c}{f_w}x + \frac{c}{Z}dx\right), \qquad \theta_R = \arctan\left(\frac{c}{f_w}x - \frac{c}{Z}dx\right), \tag{16}$$

$$\theta_y = \arctan\left(\frac{c}{f_w}y + \frac{c}{Z}dy\right), \tag{17}$$

where c is the camera conversion factor from pixel to meters; dx, dy are the terms $dx_L = dx_R$ and $dy_L = dy_R$ in pixel units.

Bearing in mind the wide-to-active camera mapping equation, in the following section we will describe the algorithm to move the active cameras to gaze and fixate in depth any object in the field of view of the wide-angle camera.

4. Fixation in depth

Two different eyes movements can be distinguished: version movements rotate the two eyes by an equal magnitude in the same direction, whereas vergence movements rotate the two eyes in opposite direction. The vergence angle, together with version and tilt angles, uniquely describe the fixation point in the 3D space according to the Donders' law (Donders, 1969).

Fixation in depth is the coordinated eye movement to align the two retinal images in the respective foveas. Binocular depth perception has its highest resolution in the well-known Panum area, i.e. a rather small area centred on the point of fixation (Kuon & Rose, 2006). The fixation of a single point in the scene can be achieved, mainly, by vergence eye-movements which are driven by binocular disparity (Rashbass & Westheimer, 1961). It follows that the amount of disparity around the Panum area must be reduced in order to properly align the two retinal images in the respective foveas.

4.1 Defining the *Panum* area

The Panum area is normally set around the centre of uncalibrated images. This particular assumption becomes a problem in systems where the images are captured by using variable-focal-length lenses; consequently, if the centre of the image is not lying on the optical axis, then any change in the field of view will produce a misalignment of the Panum area after a fixation in depth. Lenz & Tsai (1988) were the first in proposing a calibration method to determine the image centre by changing the focal length even though no zoom lenses were available at that time. In a subsequent work (Lavest et al., 1993) have used variable-focal-length lenses for three-dimensional reconstruction and they tested the calibration method proposed by (Lenz & Tsai, 1988).

In a perspective projection geometry the parallel lines, not parallel to the image plane, appear to converge to a unique point as in the case of the two verges of a road which appear to converge in the distance; this point is known as the vanishing point. Lavest et al. (1993) used the properties of the vanishing point to demonstrate that, with a zoom lens, it is possible to estimate the intersection of the optical axis and the image plane, i.e. the image centre.

The Equation 18 is the parametric representation of a set of parallel lines defined by the direction vector $\vec{D} = (D_1, D_2, D_3)$ and parameter $t \in [-\infty, +\infty]$. The vanishing point of these parallel lines can be estimated by using the perspective projection as shown in Equation 19:

$$\begin{aligned} X(t) &= X(0) + D_1 t, \\ Y(t) &= Y(0) + D_2 t, \\ Z(t) &= Z(0) + D_3 t. \end{aligned} \tag{18}$$

$$\begin{aligned} x &= \lim_{t\to\infty} f\frac{X(t)}{Z(t)} = f\frac{D_1}{D_3}, \\ y &= \lim_{t\to\infty} f\frac{Y(t)}{Z(t)} = f\frac{D_2}{D_3}, \\ z &= \lim_{t\to\infty} f\frac{Z(t)}{Z(t)} = f. \end{aligned} \tag{19}$$

The result shown in Equation 19 demonstrates that the line passing through the optical centre of the camera and the projection of the vanishing point of the parallel lines is collinear to the director vector ($\vec{D}$) of these lines as shown below:

$$\begin{bmatrix} x \\ y \\ f \end{bmatrix} = \frac{f}{D_3} \begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix}. \tag{20}$$

According to the aforementioned equations and taking into account that, by convention, the centre of the image is the intersection of the optical axis and the image plane; it is possible to conclude that the vanishing point of a set of lines parallel to the optical axis lies in the image centre. The optical zoom can be considered as a virtual movement of the scene throughout the optical axis; in this regard, any point in the scene follows a virtual line parallel to the

optical axis. This suggests that, from the tracing of two points across a set of zoomed images, it is possible to define the lines $L1$ and $L2$ (see Fig. 4) which represent the projection of these virtual lines in the image plane. It follows that the intersection of $L1$ and $L2$ corresponds with the image centre.

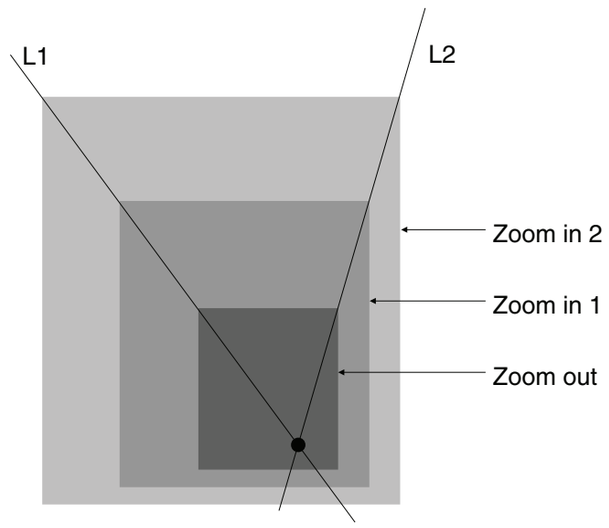


Fig. 4. Geometric determination of the image centre by using zoomed images. The intersection of the lines $L1$ and $L2$, defined by the tracing of two points across the zoomed images, corresponds with the image centre.

Once the equations of lines $L1$ and $L2$ have been estimated, it is possible to compute their intersection. Now, the Panum area is defined as a small neighbourhood around the intersection of these lines and thus it is possible to guarantee the fixation of any object even under changes in the field of view of the active cameras.

4.2 Developing the fixation-in-depth algorithm

Once the Panum area is properly defined, it is possible to develop an iterative angular-position control based on disparity estimations to fixate in depth any point in the field of view of the wide-angle camera. Fig. 5 shows a scheme of the angular-position control of the three-camera system. Any salient feature in the cyclopean image (wide-angle image) provides the point (x, y), in image coordinate, in order to set the version movement. Once the version movement is completed, the disparity estimation module can provide information about the depth of the object in the scene; this information is used to iteratively improve the alignment of the images in the active cameras.

Considering that the angular position of the cameras is known at every moment, it is possible to use the disparity information around the Panum area to approximate the scene depth; this is, a new Z in the wide-to-active camera mapping equations (see Equation 16). If we take the left image as reference, then the disparity information tells us how displaced the right image is; hence, the mean value of these disparities around the Panum area can be used to estimate the angular displacement needed to align the left and right images. As the focal length of the

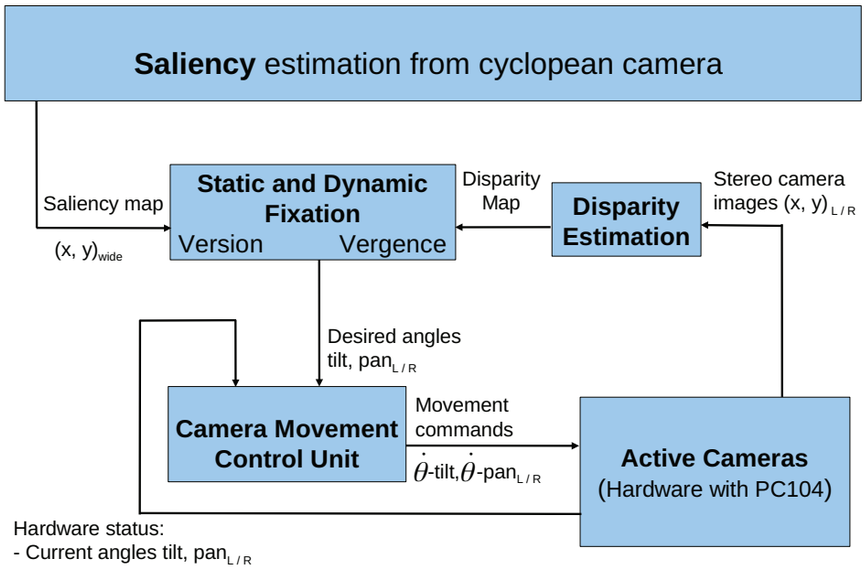


Fig. 5. Angular-position control scheme of the trinocular system.

active cameras can be approximated from the current zoom value, the angular displacement θ can be estimated as follow:

$$\theta = \arctan\left(\frac{cdx}{f}\right). \tag{21}$$

Once the angular displacement is estimated, the new Z parameter is obtained according to Equation 22:

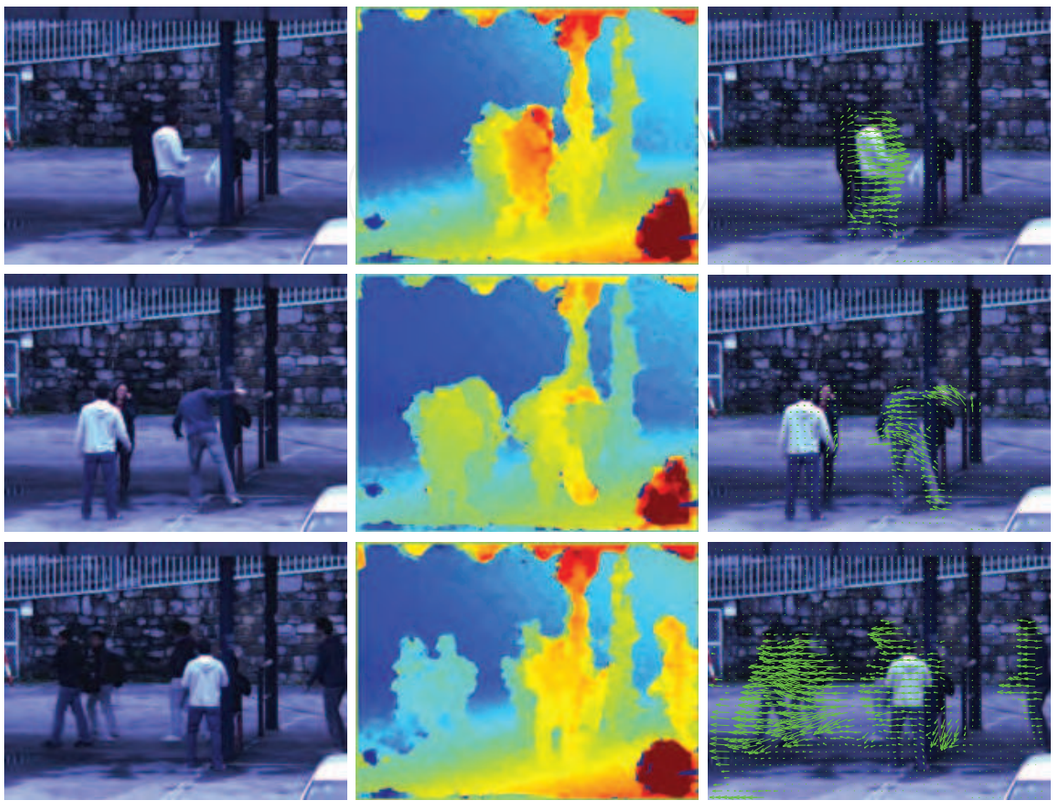
$$Z = \frac{f_w cdx}{f_w \tan(\theta_L + \theta_{verg}) - cx}. \tag{22}$$

The angle θ_{verg} is half of the angular displacement θ according to (Rashbass & Westheimer, 1961). In order to iteratively improve the alignment of the images in the active cameras, the angle θ_{verg} is multiplied by a constant ($q < 1$) in the angular-position control algorithm; this constant defines the velocity of convergence of the iterative algorithm.

5. Benefits of using binocular disparity and optic flow in image segmentation

The image segmentation is an open research area in computer vision. The problem of properly segment an image has been widely studied and several algorithms have been proposed for different practical applications in the last three decades. The perception of what is happening in an image can be thought of as the ability for detecting many classes of patterns and statistically significant arrangements of image elements. Lowe (1984) suggests that human perception is mainly a hierarchical process in which prior knowledge of the world is used to provide higher-level structures, and these ones, in their turn, can be further combined to yield new hierarchical structures; this line of thoughts was followed in (Shi & Malik, 2000). It is worth noting that the low-level visual features like motion and disparity (see Fig. 6) can offer a first description of the world in certain practical application (cf. Harville, 2004; Kolmogorov et al., 2005; Yilmaz et al., 2006; Zhao & Thorpe, 2000). The purpose of this section is to show

the benefits of using binocular disparity and optic flow estimates in segmenting surveillance video sequences rather than to make a contribution to the solution of the general problem of image segmentation.



(a) World. (b) Depth map. (c) Motion map.

Fig. 6. Example of how different scenes can be described by using our framework. The low-level visual features refer to both disparity and optic flow estimates.

The following is a case of study in which the proposed system is capable of segmenting all individuals in a scene by using binocular disparity and optic flow. In a first stage of processing, the system fixates in depth the individuals according to the aforementioned algorithm (see section 4); that is, an initial fast movement of the cameras (version) triggered by a saliency in the wide-angle camera, and a subsequent slower movement of the cameras (vergence) guided by the binocular disparity. In a second stage of processing, the system changes the field of view of the active cameras in order to magnified the region of interest. Finally, in the last stage of processing, the system segments the individuals in the scene by using a threshold in the disparity information (around disparity zero or point of fixation) and a threshold in the orientation of the optic flow vectors. The results of applying the above mentioned processing stages are shown in Fig. 7. Good segmentation results can be achieved from the disparity measures by defining a set of thresholds (see Fig. 7b), however, a better data segmentation is obtained by combining the partial segments of binocular disparity and optic flow, respectively; an example is shown in Fig. 7c.

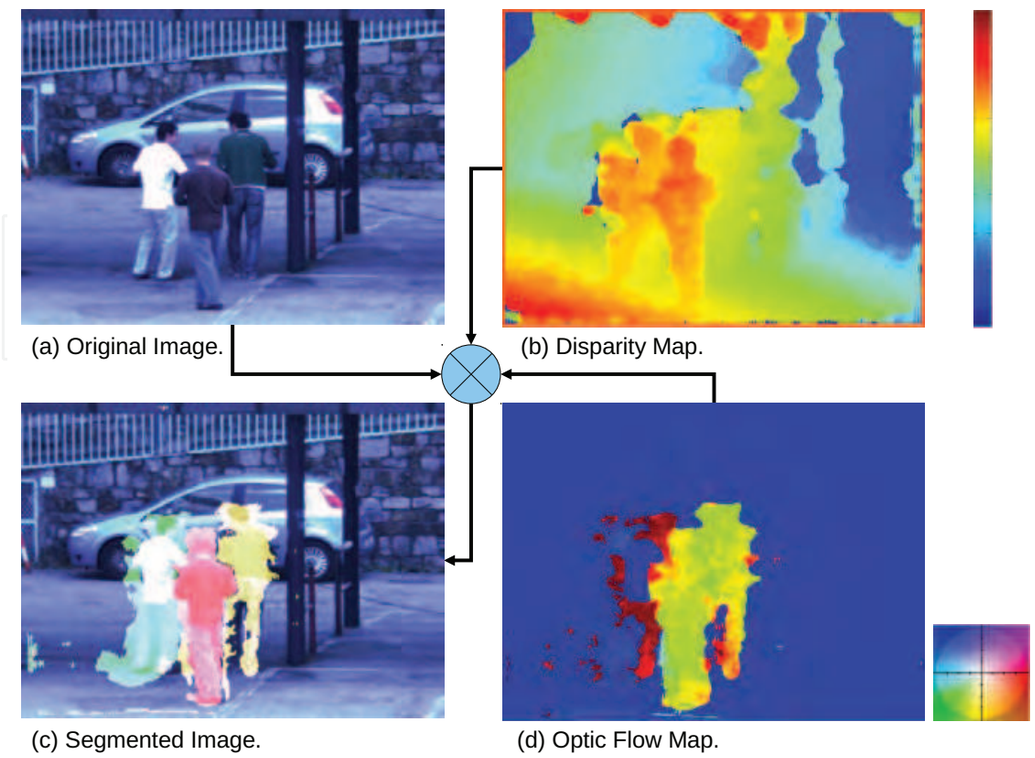


Fig. 7. Case of study: the segmentation of an image by using disparity and optic flow estimates.

The results in segmentation are constrained by the estimates of disparity and optic flow. For this reason, it is necessary to follow segmentation strategies, like the one proposed by Shi & Malik (2000), in order to achieve the appropriate robustness in the data segmentation. In fact, they argue the necessity of combining different features like colour, edge or in general any kind of texture information to create a hierarchical partition of the image, based on graph theory, in which prior knowledge is used to confirm current grouping or to guide further classifications.

6. The system performance

So far, we have presented an active vision system capable of estimating both optic flow and binocular disparity through a biologically inspired strategy, and capable of using these information to change the viewpoint of the cameras in an open, uncontrolled environment. This capability lets the system interact with the environment to perform video surveillance tasks. The purpose of this work was to introduce a novel system architecture for an active vision system rather than to present a framework for performing specific surveillance tasks. Under this perspective, it was first described the low-level vision approach for optic flow and binocular disparity, and then it was presented a robotic head which uses this approach to effectively solve the problem of fixation in depth.

In order to evaluate the performance of the system, it is necessary to differentiate the framework instances according to their role in the system. On the one hand, both optic flow and binocular disparity are to be used as prominent features for segmentation; hence, it is important to evaluate the accuracy of the proposed algorithms by using test sequences for which ground truth is available (see *http://vision.middlebury.edu/*). On the other hand, we must evaluate the system performance in relation to the accuracy of the binocular system to correctly change the viewpoint of the cameras.

6.1 Accuracy of the distributed population code

The accuracy of the estimates has been evaluated for a system with $N = 16$ oriented filters, each tuned to $M = 3$ different velocities and to $K = 9$ binocular phase differences. The used Gabor filters have a spatiotemporal support of $(11 \times 11) \times 7$ pixels $\times$ frames and are characterised by a bandwidth of 0.833 octave and spatial frequency $\omega_0 = 0.5\pi$. The Table 3 shows the results for *distributed population code* that has been applied to the most frequently used test sequences. The optic flow was evaluated by using the database described in (Baker et al., 2007) and the disparity was evaluated by using the one described in (Scharstein & Szeliski, 2002); however, in the case of disparity test sequences, the ground truth contains horizontal disparities, only; for this reason, it was also used the data set described in (Chessa, Solari & Sabatini, 2009) to benchmark the 2D-disparity measures (horizontal and vertical).

Distributed population code			
Sequences	Venus	Teddy	Cones
Disparity (%BP)	4.5	11.7	6.4
Sequences	Yosemite	Rubberwhale	Hydrangea
Optic Flow (AAE)	3.19	8.01	5.79

Table 3. Performance of the proposed distributed population code. On the one hand, the reliability of disparity measures has been computed in terms of percentage of bad pixels (%BP) for non-occluded regions. On the other hand, the reliability of optic flow measures has been computed by using the average angular error (AAE) proposed by Barron (Barron et al., 1994).

A quantitative comparison between the proposed *distributed population code* and some of the well-established algorithms in literature has been performed in (Chessa, Sabatini & Solari, 2009). The performances of the stereo and motion modules are shown in Table 3, which substantiates the feasibility of binocular disparity and optic flow estimates for image segmentation; the visual results are shown in Fig. 7.

6.2 Behaviour of the trinocular system

A good perception of the scene's depth is required to properly change the viewpoint of a binocular system. The previous results for disparity estimation have shown to be a valuable cue for 3D perception. The purpose now is to demonstrate the capability of the trinocular head to fixate any object in the field of view of the wide-angle camera. In order to evaluate the fixation in depth algorithm, two different scenarios have been considered: the long-range scenario in which the depth is larger than 50 meters in the line of sight (see Fig. 8), and the short-range scenario in which the depth is in the range between 10 and 50 meters (see Fig. 11).

(a) Cyclopean Image.

(b) Left Image, point A.

(c) Right Image, point A.

(d) Left Image, point B.

(e) Right Image, point B.

(f) Left Image, point C.

(g) Right Image, point C.

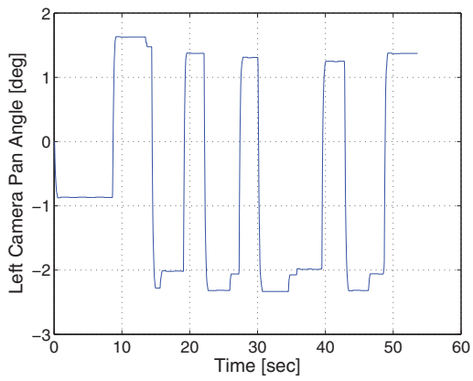
Fig. 8. Long-range scenario: Fixation of points A, B and C. A zoom factor of 16x was used in the active cameras. Along the line of sight the measured depths were approximately 80 m, 920 m, and 92 m, respectively.

The angular-position control uses the disparity information to align the binocular images in the Panum area. In order to save computational resources and considering that just a small area around the centre of the image has the disparity information of the target object, the size of the Panum area has been empirically chosen as a square region of 40x40 pixels. Accordingly, the mean value of the disparity in the Panum area is used to iteratively estimate the new Z parameter.

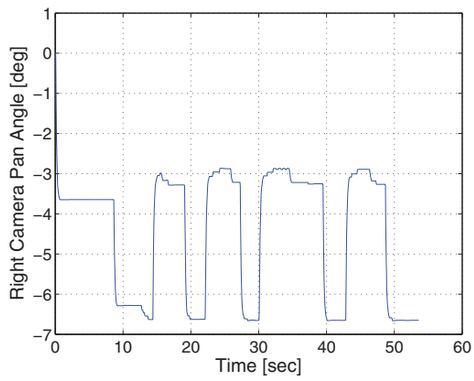
In order to evaluate the performance of the trinocular head, we first tested the fixation strategy in the long-range scenario. In the performed tests, three points were chosen in the cyclopean image (see Fig. 8(a)). For each point, the active cameras performed a version movement according to the coordinate system of the cyclopean image and, inmediately after, the angular-position control started the alignment of the images by changing the pan angles iteratively. Once the images were aligned, a new point in the cyclopean image was provided.

Fig. 9 shows the angular changes of the active cameras during the test in the long-range scenario. In Figs. 9(a) and 9(b) the pan angle of the left and right cameras, respectively, is depicted as a function of time. Fig. 9(c) shows the same variation for the common tilt angle. Each test point of the cyclopean image was manually selected after the fixation in depth of the previous one; consequently, the plots show the angular-position control behaviour during changes in the viewpoint of the binocular system. It is worth noting that the version

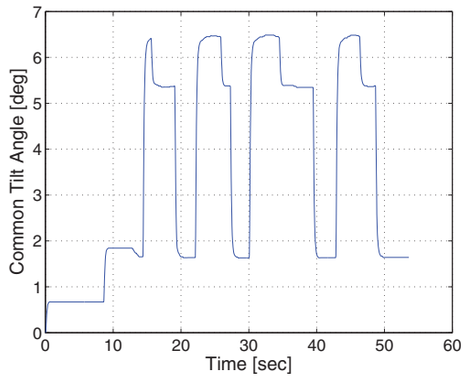
movements correspond, roughly speaking, with the pronounced slopes in the graphs, while the vergence movements are smoother and therefore with a less pronounced slope.



(a) Left camera pan movements.



(b) Right camera pan movements.

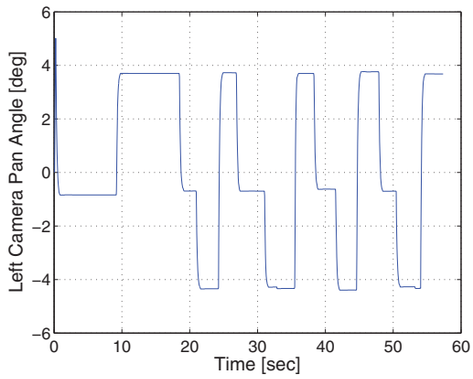


(c) Common tilt movements.

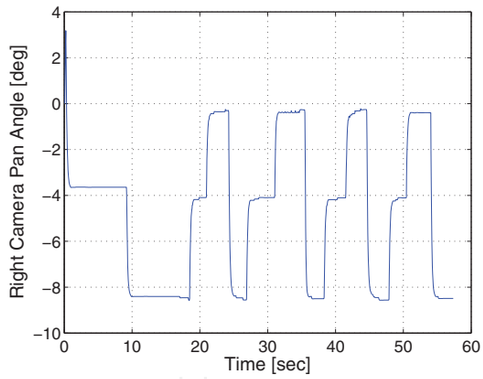
Fig. 9. Temporal changes in the angular position of the active cameras to fixate in depth the points A, B and C in a long-range scenario.

In a similar way, the fixation in depth algorithm was also evaluated in short-range scenarios by using three test points (see Fig. 11). We followed the same procedure used for long-range scenarios and the results are shown in Fig. 10.

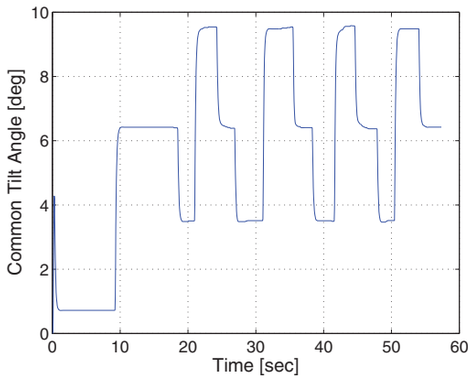
From the plots in Figs. 9 and 10 we can observe that small angular shifts were performed just after a version movement; this behaviour is due to two factors: (1) the inverse relationship between the vergence angle and the depth by which for large distances the optical axes of the binocular system can be well approximated as parallel; and (2) the appropriate geometrical description of the system which allows us to properly map the angular position of the active cameras with respect to the cyclopean image. Actually, there are not enough differences between long and short-range scenarios in the angular-position control, because the vergence angles begin to be considerable for depths minor than 10 meters, approximately; it is worth noting that, this value is highly dependent on the baseline of the binocular system.



(a) Left camera pan movements.



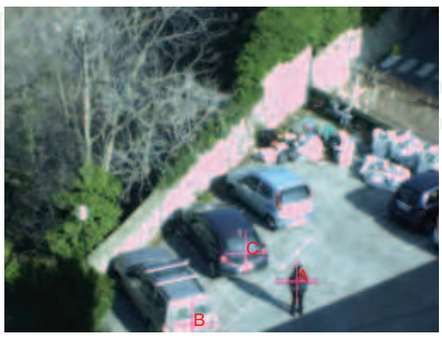
(b) Right Camera pan movements.



(c) Common tilt movements.

Fig. 10. Temporal changes in the angular position of the active cameras to fixate in depth the points A, B and C in a short-range scenario.

Finally, the justification for using two different scenarios is the field of view of the active cameras. Even though the wide-to-active camera mapping equations do not depend on the field of view of the active cameras, everything else does. It follows that the estimation of optic flow and disparity loses resolution due to narrow fields of view in the active cameras. In order to clarify the system behaviour, it is worth to highlight that the framework always performs the fixation in depth by using the maximum field of view in the active cameras, and immediately after, it changes the field of view of the cameras according to the necessary magnification. In this regard, the adequate definition of the Panum area plays an important role in the framework (see section 4.1). Consequently, Figs. 8 and 11 show the performance of the framework not only in terms of the fixation but also for a proper synchronisation of all processing stages in the system; these images were directly obtained from the system during the experiments in Figs. 9 and 10. Fig. 8 shows the fixation in depth of three test points. The zoom factor of the active cameras in all cases was 16x. The angular-position control estimated the depth along the line of sight for each fixated target and the approximated values were 80 m, 920 m, and 92 m, respectively. Likewise, Fig. 11 shows the fixation in depth of three test points at different zoom factors each one, namely: 4x, 16x, and 4x, respectively. Along the line



(a) Cyclopean Image.

(b) Left Image, point A.

(c) Right Image, point A.

(d) Left Image, point B.

(e) Right Image, point B.

(f) Left Image, point C.

(g) Right Image, point C.

Fig. 11. Short-range scenario: Fixation of points A, B, and C. The different zoom factors used in the active cameras were 4x, 16x, and 4x, respectively. Along the line of sight the measured depths were approximately 25 m, 27 m, and 28 m, respectively.

of sight the measured depths were approximately 25 m, 27 m, and 28 m, for points A, B, and C, respectively.

7. Conclusion

We have described a trinocular active visual framework for video surveillance applications. The framework is able to change the viewpoint of the active cameras toward areas of interest, to fixate a target object at different fields of view, and to follow its motion. This behaviour is possible thanks to a rapid angular-position control of the cameras for object fixation and pursuit based on disparity information. The framework is capable of recording image frames at different scales by zooming individual areas of interest, in this sense, it is possible to exhibit the target's identity or actions in detail. The proposed visual system is a cognitive model of visual processing replicating computational strategies supported by the neurophysiological studies of the mammalian visual cortex which provide the system with a powerful framework to characterise and to recognise the environment, in this sense, the optic flow and binocular disparity information are an effective, low-level, visual representation of the scenes which provide a workable base for segmenting the dynamic scenarios; it is worth noting that, these measures can easily disambiguate occlusions in the different scenarios.

8. References

Adelson, E. & Bergen, J. (1985). Spatiotemporal energy models for the perception of motion, *JOSA* 2: 284–321.

Adelson, E. & Bergen, J. (1991). The plenoptic and the elements of early vision, *in* M. Landy & J. Movshon (eds), *Computational Models of Visual Processing*, MIT Press, pp. 3–20.

Andrade, E. L., Blunsden, S. & Fisher, R. B. (2006). Hidden markov models for optical flow analysis in crowds, *Pattern Recognition, International Conference on* 1: 460–463.

Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M. & Szeliski, R. (2007). A database and evaluation methodology for optical flow, *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1 –8.

Barron, J., Fleet, D. & Beauchemin, S. (1994). Performance of optical flow techniques, *Int. J. of Computer Vision* 12: 43–77.

Chessa, M., Sabatini, S. & Solari, F. (2009). A fast joint bioinspired algorithm for optic flow and two-dimensional disparity estimation, *in* M. Fritz, B. Schiele & J. Piater (eds), *Computer Vision Systems*, Vol. 5815 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 184–193.

Chessa, M., Solari, F. & Sabatini, S. (2009). A virtual reality simulator for active stereo vision systems, *VISAPP* .

Daugman, J. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, *JOSA* A/2: 1160–1169.

Donders, F. C. (1969). Over de snelheid van psychische processen, *Onderzoekingen gedann in het Psychologish Laboratorium der Utrechtsche Hoogeschool: 1868-1869. Tweede Reeks, II, 92-120.*, W. E. Koster (Ed.) and W. G. Koster (Trans.), pp. 412 – 431. (Original work published 1868).

Fleet, D., Wagner, H. & Heeger, D. (1996). Neural encoding of binocular disparity: Energy models, position shifts and phase shifts, *Vision Res.* 36(12): 1839–1857.

Hansard, M. & Horaud, R. (2008). Cyclopean geometry of binocular vision, *Joural of the Optical Society of America A* 25(9): 2357–2369.

Harville, M. (2004). Stereo person tracking with adaptive plan-view templates of height and occupancy statistics, *Image and Vision Computing* 22(2): 127 – 142. Statistical Methods in Video Processing.

Helmholtz, H. v. (1925). *Treatise on Physiological Optics*, Vol. III, transl. from the 3rd german edn, The Optical Society of America, New York, USA.

Hubel, D. H. & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex, *The Journal of Physiology* 195(1): 215–243.

Kapur, J., Sahoo, P. & Wong, A. (1985). A new method for gray-level picture thresholding using the entropy of the histogram, *Computer Vision, Graphics, and Image Processing* 29(3): 273 – 285.

Koenderink, J. & van Doorn, A. (1976). Geometry of binocular vision and a model for stereopsis, *Biological Cybernetics* 21: 29–35.

Kolmogorov, V., Criminisi, A., Blake, A., Cross, G. & Rother, C. (2005). Bi-layer segmentation of binocular stereo video, *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* 2: 407–414.

Kumar, R. K., Ilie, A., Frahm, J.-M. & Pollefeys, M. (2008). Simple calibration of non-overlapping cameras with a mirror, *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* 0: 1–7.

Kuon, I. & Rose, J. (2006). Measuring the gap between fpgas and asics, *FPGA '06: Proceedings of the 2006 ACM/SIGDA 14th international symposium on Field programmable gate arrays*, ACM, New York, NY, USA, pp. 21–30.

Lavest, J.-M., Rives, G. & Dhome, M. (1993). Three-dimensional reconstruction by zooming, *Robotics and Automation, IEEE Transactions on* 9(2): 196–207.

Lee, L., Romano, R. & Stein, G. (2000). Monitoring activities from multiple video streams: establishing a common coordinate frame, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22(8): 758 –767.

Lenz, R. & Tsai, R. (1988). Techniques for calibration of the scale factor and image center for high accuracy 3-d machine vision metrology, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10: 713–720.

Lowe, D. G. (1984). *Perceptual Organization and Visual Recognition*, PhD thesis, STANFORD UNIV CA DEPT OF COMPUTER SCIENCE.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Henry Holt and Co., Inc., New York, NY, USA.

Ohzawa, I., DeAngelis, G. & Freeman, R. (1990). Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors, *Science* 249: 1037–1041.

Otsu, N. (1979). A threshold selection method from graylevel histograms, *IEEE Trans. Syst., Madn. & Cybern.* 9: 62–66.

Pauwels, K. & Van Hulle, M. M. (2006). Optic flow from unstable sequences containing unconstrained scenes through local velocity constancy maximization, *British Machine Vision Conference (BMVC 2006)*, Edinburgh, Scotland, pp. 397–406.

Pouget, A., Dayan, P. & Zemel, R. S. (2003). Inference and computation with population codes., *Ann. Rev Neurosci* 26: 381–410.

Rashbass, C. & Westheimer, G. (1961). Disjunctive eye movements, *The Journal of Physiology* 159: 339–360.

Ratha, N. & Jain, A. (1999). Computer vision algorithms on reconfigurable logic arrays, *Parallel and Distributed Systems, IEEE Transactions on* 10(1): 29 –43.

Ridler, T. W. & Calvar, S. (1978). Picture thresholding using an iterative selection method, *Systems, Man and Cybernetics, IEEE Transactions on* 8(8): 630 –632.

Scharstein, D. & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, *Int. J. of Computer Vision* 47: 7–42.

Shi, J. & Malik, J. (2000). Normalized cuts and image segmentation, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22(8): 888 –905.

Stauffer, C. & Grimson, W. (1999). Adaptive background mixture models for real-time tracking, *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, Vol. 2, pp. 2 vol. (xxiii+637+663).

Stauffer, C. & Grimson, W. (2000). Learning patterns of activity using real-time tracking, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22(8): 747 –757.

Theimer, W. & Mallot, H. (1994). Phase-based binocular vergence control and depth reconstruction using active vision, *CVGIP: Image Understanding* 60(3): 343–358.

Tsai, R. (1987). A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses, *Robotics and Automation, IEEE Journal of* 3(4): 323 –344.

Weems, C. (1991). Architectural requirements of image understanding with respect to parallel processing, *Proceedings of the IEEE* 79(4): 537 –547.

Yilmaz, A., Javed, O. & Shah, M. (2006). Object tracking: A survey, *ACM Comput. Surv.* 38.

Zhao, L. & Thorpe, C. (2000). Stereo- and neural network-based pedestrian detection, *Intelligent Transportation Systems, IEEE Transactions on* 01(3): 148 –154.

Machine Vision - Applications and Systems

Edited by Dr. Fabio Solari

ISBN 978-953-51-0373-8

Hard cover, 272 pages

Publisher InTech

Published online 23, March, 2012

Published in print edition March, 2012

Vision plays a fundamental role for living beings by allowing them to interact with the environment in an effective and efficient way. The ultimate goal of Machine Vision is to endow artificial systems with adequate capabilities to cope with not a priori predetermined situations. To this end, we have to take into account the computing constraints of the hosting architectures and the specifications of the tasks to be accomplished, to continuously adapt and optimize the visual processing techniques. Nevertheless, by exploiting the low?cost computational power of off?the?shell computing devices, Machine Vision is not limited any more to industrial environments, where situations and tasks are simplified and very specific, but it is now pervasive to support system solutions of everyday life problems.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen